DOCUMENT RESUME

ED 042 794 TM 000 052

AUTHOR Pinsky, Paul; Gorth, William P.

TITLE Descriptive Analysis of HS420: Eleventh Grade Algebra, First Semester.

INSTITUTION Massachusetts Univ., Amherst. School of Education.
SPONS AGENCY Charles F. Kettering Foundation, Dayton, Ohio.
REPORT NO TM-21
PUB DATE Jul 69
NOTE 22p.

EDRS PRICE EDRS Price MF-$0.25 HC-$1.20
DESCRIPTORS Achievement Tests, *Algebra, Cognitive Tests, Course Objectives, Grade 11, Measurement, *Measurement Instruments, *Measurement Techniques, *Predictive Ability (Testing), Predictor Variables, *Test Construction
IDENTIFIERS CAM, *Comprehensive Achievement Monitoring

ABSTRACT

Analysis of data on this algebra course, gathered by the Comprehensive Achievement Monitoring (CAM) system, indicated that equivalent scores were yielded by either random or chronological arrangement of the test items on the monitor forms. Chronological arrangement may be permissible, therefore, for normal data processing; random arrangement should be retained for computer processing. A set of nine cognitive ability tests were found to be poor predictors of achievement and, hence, ineffective for scheduling students. Split-half reliability for each text administration and test-retest reliability for each pair of administrations are given, together with test difficulties. An attempt was made to fit learning curves to the data to provide measures of individual performance. No consistent pattern appeared as to the curve providing the best fit over all students and it is suggested that meaningful information cannot be gathered from comprehensive monitors containing only nine items, as did these. The anticipated increase in scores from pretests to posttests did not occur. Since reliabilities and other parameters of the tests were acceptable, other explanations must be sought for the lack of change. For example, content validity of items or relatively sophisticated prior knowledge held by the students. (DG)

Project Comprehensive
Achievement
Monitoring

Technical Memorandum No. TM-21 July 1969

DESCRIPTIVE ANALYSIS OF HS420: ELEVENTH GRADE ALGEBRA, FIRST SEMESTER

by

P. Pinsky and W. Gorth
Stanford University

The research and development reported herein was performed pursuant to a grant from the Charles F. Kettering Foundation to the Principal Investigator, Dr. Dwight W. Allen, Dean, School of Education, The University of Massachusetts. The Project CAM staff includes D. Evans, W. Gorth, P. Pinsky, N. Sims, L. Wightman, and G. Worle.

Additional information or permission to quote from this document or to reproduce it, wholly or in part, should be obtained from:

William P. Gorth
Project CAM
School of Education
Stanford University
Stanford, California 94305

or

David R. Evans
Project CAM
School of Education
The University of Massachusetts
Amherst, Massachusetts 01002

Descriptive Analysis of HS420
Eleventh Grade Algebra, First Semester

by
Paul Pinsky and William Gorth
Stanford University

This analysis is of the course HS420, which used CAM monitoring during the 1968-1969 school year. The course is the first semester of a two-semester eleventh grade mathematics course which was taught by the traditional, teacher-paced method. The analysis is descriptive because the data did not behave as expected under a CAM model, i.e., they did not show an increase in student achievement during the school year.

The data collection was excellent and results were returned to the students two or three days after monitoring. The CAM monitors were the only tests used. The reliabilities and standard errors of all tests used, including the CAM pretest and posttest, were in the range expected and the distribution of the students' criterion scores (Lindeman, Gorth, & Allen, 1968) was of a truncated normal form.

The analyses indicated the following:

1. Random versus chronological arrangements of items on the monitor forms yield equivalent scores; therefore, a chronological arrangement of items is probably permissible for manual data processing. However, due to the lack of increases in achievement, some doubt concerning the effect of arrangement still exists. Items should be randomly arranged when a computer is used for the data processing.

2. Students' criterion scores should not be compared under the CAM model presently used.

3. Cognitive ability test scores (TM-18) did not appear very useful in scheduling students to take various monitor forms througout

the year; i.e., the scores were poor predictors of achievement. The results indicate that in scheduling students, their grades in similar courses should be used rather than the current battery of cognitive ability tests.

4. The change in students' scores throughout the semester did not behave as we expected. Scores on the pretests were much higher than we anticipated, averaging over 40%, while the posttest averaged only slightly higher at about 60%. The reliabilities and other statistical parameters of the tests were acceptable. Therefore, other explanations for the lack of change in achievement, e.g., content validity of items relative to the course or relatively sophisticated prior knowledge by students, must be explored.

Test Schedules

Fourteen different sets of nine items were used. However, the nine items are arranged differently, once randomly and once chronologically, in the order of presentation of the content they measured, yielding 28 distinct test forms. One 18-item pretest and one 36-item posttest were used. This is the first year in the Project CAM for HS420. No item analysis was available for the items.

Fourteen weekly testing periods were used during the semester. Each student will have two forms which he will take three times in a row, two forms which he will take twice in a row, and four forms which he will see once. There are 14 testing schedules, each of which uses the test forms with the items arranged either randomly or chronologically. Therefore we have a total of 28 distinct testing schedules. Either four or five students are in each testing schedule. The scheduling was based on the pretest data; i.e., 18 items randomly chosen from the same item pool used for the weekly monitoring. The scores ranged from 1 correct to 15 correct with a median of 8 correct. The students were divided into groups scoring low, medium, or high on the pretest and the scheduling was completed by assigning an equal number of each of the three student

groups to each of the various test schedules. (See TM-17 for schedule groups.)

Scheduling Procedures

In Project CAM, students are scheduled to take monitor forms using stratified random sampling. Therefore, a representative sample of the students take each monitor form at each test administration. However, the stratification in HS420 was based on a single CAM pretest form score. The students scored quite high on the pretest. The correlation between the single pretest and the single posttest wes .604. In most CAM courses we would hope that the course would be designed so that the students would score low on the pretest. Therefore, the pretest would not be a good predictor of a posttest performance. Although here the pretest was a good criterion for stratification, generally one would not expect this result. When one reflects upon the goals of the CAM monitoring system, a single CAM pretest form seems to have little value.

Nine reference tests of cognitive ability (TM-18 and French, et al, 1963) were administered during the semester and their potential for stratifying students was investigated. A correlation of the nine ability scores with the posttest score reveals their poor predictive power in this course (see Table TM-21).

.....

INSERT TABLE TM-21.1 ABOUT HERE

.....

In addition, a step-wise regression was used to predict the final test score from these nine ability covariates. Seven of the nine covariates (the last two being non-significant) yielded a multiple correlation coefficient of .215.

For HS420, a pretest of course content was a more suitable criterion for stratification then the cognitive ability tests which were used.

Table TM-31.1 Correlation of Posttest of Achievement with Cognitive Ability Test Scores

Number[a]	Test name[b]	Correlation
1	Wide Range Vocabulary	.047
2	Number Comparison	.104
3	Surface Development	.043
4	Cube Comparison	.015
5	Letter Sets	.147
6	Word Arrangement	.004
7	Inference	.162
8	Maze Tracing	.105
9	Auditory number Span	.061

Note.-- N = 107

[a] Tests numbered in the order in which they were administered.

[b] Tests taken from French, et al (1963) and are described in TM-10.

Monitor Form Characteristics

It was initially hoped that the CAM project would not only provide information about group performances on the various performance criteria in the course but also about individual students. However, our analyses have indicated that any information about individual students and differences between individual students at each test administration should not be attempted using the total score on CAM monitors.

Some analyses were performed on the characteristics of the total score on various monitor forms. We investigated the equivalence of the difficulty level of the various test forms. Because we had 28 monitor forms and only 140 students, this gave only four or five students taking a monitor form at each test administration. It was decided that this number of observations per administration was not sufficient to estimate average test difficulty. Therefore, scores on the same test administered at consecutive times were grouped to provide a more stable estimate. Scores from test administrations 2, 3, and 4 were grouped in section 1; administrations 5, 6, 7, and 8 were grouped into section 2; and administrations 9, 11, 12, and 13 were grouped into section 3 (tests were not administered at time 10 due to a clerical error). Table TM-21.2 presents the average number correct on each of the test forms for each of these sections. Remembering that there are only fourteen different sets of items and that each set of items was arranged in two ways, the items, arranged in two different fashions, are presented in the same row, enabling one to look for differences between the random and chronological arrangements.

INSERT TABLE TM-21.2 ABOUT HERE

Another question is whether the difficulty changes over time. Table TM-21.3 indicates if a test form was easier than the grand mean test form during that section by a plus; if it was more difficult by a minus. Changes in test form difficulty over time would be indicated by minus forms during section 1 changing to plusses during section 3 or

Table TM-21.2 Average Number Correct by Form Over Time

Test form[a]		Section 1[a]		Section 2[a]		Section 3[a]		Mean across all sections[a]	
Random[b]	Chronological[b]	Rand	Chron	Rand	Chron	Rand	Chron	Rand	Chron
2	11	5.9	5.8	6.1	6.2	6.2	6.4	6.1	6.1
7	6	5.5	4.8	6.1	5.9	5.6	5.8	5.7	5.5
9	15	5.6	5.9	6.5	6.2	6.8	5.2	6.3	5.8
13	24	5.2	5.3	6.1	5.6	5.9	5.5	5.7	5.5
14	4	6.7	6.1	6.1	6.0	7.0	6.3	6.6	6.1
16	29	6.2	6.4	5.4	6.2	6.6	6.3	6.1	6.3
17	23	6.9	5.7	6.5	5.5	6.9	6.2	6.5	5.8
19	28	5.2	5.1	4.2	5.2	5.7	4.8	5.0	5.0
20	18	7.0	6.7	7.4	7.2	7.2	6.9	7.2	6.9
21	30	5.1	3.7	5.3	4.6	5.8	6.6	5.4	5.0
22	8	5.2	5.3	5.2	5.4	5.6	5.1	5.3	5.3
25	10	6.0	5.7	6.5	6.3	7.6	5.8	6.7	6.0
26	3	3.8	4.9	6.5	4.2	6.1	5.7	5.5	4.9
27	5	4.6	4.5	5.8	4.7	6.0	4.9	5.5	4.7
ALL FORMS		5.6	5.4	6.0	5.7	6.4	5.8	6.0	5.6

Note.-- Section 1 includes test administrations 2, 3, and 4; section 2 includes 5, 6, 7, and 8; section 3 includes 9, 11, 12, and 13.

[a] Forms are listed so that forms with identical items, arranged randomly or chronologically, are in the same row.

[b] Random and chronological refer to the arrangement of items on forms.

vice-versa. An inspection of this table indicates that little or no such change occurs.

INSERT TABLE TM-21.3 ABOUT HERE

It appears that monitor forms were not exactly equal in difficulty but that their difficulty seemed to remain constant over time.

Split-halves internal reliability were calculated at each test administration and test-retest reliabilities for each pair of administrations. These reliabilities enabled us to calculate the standard error of measure of the monitor forms for each test administration. In computing the reliability coefficients it was assumed that all the monitor forms contained the same items because they are parallel . The individual forms would not produce meaningful results due to insufficient observations. The test-retest and split-halves internal reliability are presented in Table TM-21.4.

INSERT TABLE TM-21.4 ABOUT HERE

The values in Table TM-21.4 are of the order of magnitude expected for nine-item tests (Payne and McMorris, 1967).

One of the objectives of Project CAM is to discover the effects of different arrangements of items on monitor forms. Two arrangements, random and chronological, were used in this course. Referring back to Table TM-21.2, one can see the total number correct for forms of each arrangement. Considering the standard error of measurement reported, most of the test forms appear to be of equal difficulty.

However, the grand mean of the forms with random arrangement of items is higher, 6.0, than that of the forms with the chronological items, 5.6. This difference seems to be explained by the difference in

Table TM-21.7 Individual Test Form Difficulty Related to Average Form Difficulty for Time Sections

Test form	Section 1	Section 2	Section 3	All sections
2	+	+	+	+
3	-	-	-	-
4	+	+	+	+
5	-	-	-	-
6	-	+	-	-
7	-	+	-	-
8	-	-	-	-
9	+	+	+	+
10	+	+	-	+
11	+	+	+	+
13	-	+	-	-
14	+	+	+	+
15	+	+	-	o
16	+	-	+	+
17	+	+	+	+
18	+	+	+	+
19	-	-	-	-
20	+	+	+	+
21	-	-	-	-
22	-	-	-	-
23	+	-	+	o
24	-	-	-	-
25	+	+	+	+
26	-	+	-	-
27	-	-	-	-
28	-	-	-	-
29	+	+	+	+
30	-	-	+	-

Note.-- Section 1 includes test administrations 2, 3, and 4; section 2 includes 5, 6, 7, and 8; section 3 includes 9, 11, 12, and 13.
+ indicates that the form difficulty is above the average for the corresponding section.
- indicates that the form difficulty is below the average for the corresponding section.
o indicates that the form difficulty is equal to the average for the corresponding section.

Table TM-21.4 Characteristics of Tests Across All Forms for Each Test Administration

Test administration	Test-retest reliability	Standard deviation	Standard error of measurement	Split-halves reliability
2	.48	1.80	1.30	.29
3	.43	1.71	1.29	.22
4	.50	1.87	1.32	.26
5	.67	1.80	1.04	.19
6	.46	1.91	1.41	.26
7	.41	1.75	1.34	.27
8	.38	1.72	1.36	.29
9	.41	1.85	1.42	.29
11	.51	1.79	1.26	.25
12	.55	1.74	1.17	.24
13				.27

[a] Test-retest reliability is calculated from test administration n to n+1 and is recorded in the row for test administration n.

mean score on the posttest of the students taking either only the random or only the chronological arrangement during the year. Table TM-21.5 presents the average posttest scores of students taking each monitor form at each test administration throughout the semester.

INSERT TABLE TM-21.5 ABOUT HERE

One might run a detailed analysis using the posttest as a covariate to verify statistically that there is no difference in the mean score attributable to arrangement of items.

To further search for possible differences between random and chronological arrangements of items on test forms, the split-halves internal reliability coefficients were calculated during each test administration for both the random and chronological tests (TM-21.6).

INSERT TABLE TM-21.6 ABOUT HERE

Once again there appears to be no consistent difference between the random and chronological tests.

The statistics calculated above are most relevant where the tests are designed to differentiate between students rather than predict group performances. Therefore, it is necessary to examine the effect of various arrangements of items on the tests with respect to measures of group performance. Correspondingly, achievement profiles were calculated for the two groups of students; i.e., those taking random tests and those taking chronological tests, for several different dimensions. These dimensions included (a) all questions in the course, (b) those questions related to each unit, 1, 2, 3, 4, and 5; and (c) those questions related to each topic; i.e., exponents, rational equations, inequalities, applications, and linear equations. No consistent differences can be found between the students taking random or chronological arrangements of items. A more detailed analysis might calculate

Table TM-21.5 Mean Posttest Scores of Students Taking Each Monitor Form Across All Test Administrations

Random forms	Mean posttest score	Chronological forms	Mean posttest score
2	21.9	11	20.6
7	21.0	6	20.5
9	21.9	15	21.0
13	21.6	24	19.9
1	21.6	4	19.8
16	21.5	29	20.5
17	22.8	23	21.0
19	21.9	28	19.4
20	23.2	18	21.0
21	21.0	30	20.1
22	21.8	8	20.3
25	22.1	10	20.7
26	20.8	3	19.9
27	22.0	5	20.7
Grand mean	21.8	Grand mean	20.4

Note.-- Random and chronological test forms appearing in the same row contain the same items in different arrangements.

Table TM-21.6 Split-half Reliabilities Across All Forms for Each Test Administration

Test administration	Item arrangement on forms	
	Random	Chronological
2	.27	.30
3	.16	.27
4	.34	.17
5	.23	.15
6	.35	.17
7	.33	.21
8	.43	.15
9	.20	.38
11	.22	.28
12	.32	.16
13	.32	.22

the variance in the achievement profiles and run a statistical significance test as to differences.

An analysis of the pretest and posttest was done. There was one pretest and one posttest given to all students in the class. The split-halves internal reliability coefficients were .39 for the pretest and .60 for the posttest. The standard error of measurement was 2.24 for the pretest and 3.02 for the posttest. These values are expected for an achievement test used to differentiate between individuals.

Positional Effects

Project CAM has attempted to determine whether student fatigue or warm-up effects were affecting results. The objective was to try to determine an optimal length of the monitor forms in the CAM system. The analysis considered the forms in which the items were randomly arranged and summed the total number of correct responses by position across these forms. It was performed and indicated no consistent pattern.

Repetition of Test Forms

Project CAM is interested in the effect of repeated testing upon learning. An experimental design was therefore developed, having students take the same CAM monitor form several weeks in succession. Each student was scheduled to take one test form three successive times and a different test form two successive times during the year. However, the omission of test administration 10 partially disrupted the design. The average change in total score of those students repeating the same test was calculated for three changes; i.e., the change from the first to the second administration, that from the first to the third, and that from the second to the third. Table TM-21.7 presents these results.

INSERT TABLE TM-21.7 ABOUT HERE

Table TM-21.7 Comparison of Changes in Mean Student Score for Repeated Administration of Forms

Test administration	Mean[a] score	Change in mean between administrations				
		1st & 2nd[a]	1st & 3rd[a]	1st & 2nd[b]	2nd & 3rd[b]	1st & 3rd[b]
2	5.22					
3	5.48	.26		.35		
4	5.67	.19	.45	.56	-.07	.73
5	5.76	.09	.28	.25	.90	.90
6	5.58	-.18	-.09	.21	-.35	.00
7	5.84	.26	.08	.36	.86	.71
8	6.03	.19	.45	.64	.09	.91
9	5.94	-.09	.10	.43	.22	.81
11	6.12	.18	.09	.63		
12	6.30	.18	.36	.30		
13	6.03	-.27	-.09	.30	-.25	.20

[a] Means and changes calculated for all students and all forms in per cent questions answered correctly.

[b] Changes in mean only for repeated administration of the same test forms.

The mean change in scores for tests repeated once increased by an amount greater than the average increase for all students. However, the change from the second to third administration was not as great as the change from the first to the second. Of course, the first to the third administration change in score is the sum of the first two effects. (The scores do not add up to the same change because the first to second administration includes many students who only took the test form twice in succession as well as those who took the test form three times in succession.)

In addition, the test-retest reliability was calculated for students in each period who repeated a test, Table TM-21.8.

INSERT TABLE TM-21.8 ABOUT HERE

It might be noted that these test-retest reliabilities are higher than those calculated for all students regardless of the test forms they took (see Table TM-21.4). The second to third administration and first to third administration test-retest reliabilities are not too meaningful because of a small number of observations.

Individual Differences

As mentioned previously in this report, it was hoped that measures of individual student performance could be obtained from the CAM system. We attempted fitting various learning curves to the data for all students. The BMD05R program was used to fit a first, second, and third degree curve to the total number of correct responses of each of the students for the test administrations. A subjective observation was that there appeared to be no consistent pattern as to whether a linear, quadratic, or cubic curve was the best fit over all the students.

As a further analysis to attempt to attribute some meaning to this curve fitting, a correlation was run between the following variables

Table TM-21.8 Test-retest Reliabilities for each Test Administration for Repeated Presentations of the Same Form

Test administration	Presentation		
	1st to 2nd (N=60)	2nd to 3rd (N=20)	1st to 3rd (N=20)
2	.54	-	-
3	.69	.40	.44
4	.60	.36	.00
5	.75	.80	.67
6	.68	.91	.81
7	.51	.93	.59
8	.21	.34	.45
11	.56	-	-
12	.68	.75	.16

which were calculated for each student: pretest score, posttest score, 0 to 60 day criterion score (TM-6), -200 to -10 day criterion score, a 60 to 200 day criterion score from the item analysis program, the average number of items correct over all the periods, slope of the best fit linear line of the students' data, the standard error of this slope, and the change from pretest to posttest. These correlations are presented in Table TM-21.9.

INSERT TABLE TM-21.9 ABOUT HERE

The conclusion that should be drawn from the analysis is that even using smoothing techniques such as fitting curves to the data of individual students, virtually no meaningful information can be gained about these individual students' learning curves when comprehensive monitors containing only nine items are used.

Group Performance

The group performance parameters did not follow the CAM model. A summary of the class performance on the five units (or chapters) of the course for each of the fourteen test administrations is given in Table TM-21.10.

INSERT TABLE TM-21.10 ABOUT HERE

It should be noted that on the pretest, which was given at the beginning of the semester, the class scored approximately 40% on all the material and well over 50% on the first two units. This would indicate that the class was initially quite competent in certain areas of the course which were being taught throughout the semester. Testing did not begin immediately after the pretest, so there is actually a few-week gap between the pretest and period 2. It is possible to work out an exact significance test of the change in the percentages correct in the table as presented. However, in this course it was not done, but will be presented in subsequent CAM analyses.

Table TM-21.9 Correlations of Various Measures of Student Performance (N=107)

No.	Source	2	3	4	5	6	7	8	9
1	Pretest	.60	.56	.87	.52	.56	.02	-.12	.01
2	Posttest		.73	.64	.80	.73	.17	-.11	.80
3	Criterion score: 0 to 60 days			.65	.73	.91	.13	-.13	.50
4	Criterion score: -200 to -10 days				.55	.72	-.14	-.07	.14
5	Criterion score: 60 to 200 days					.73	.22	-.15	.62
6	Mean number correct for the year						.02	-.20	.49
7	Slope of total scores on monitors across time							-.27	.20
8	Standard error of slope								-.05
9	Change from pretest to posttest								

Table TM-21.10 Percentage of Correct Responses by Unit and Test Administration

Test administration	Unit 1	2	3	4	5
Pretest	59	57	42	49	28
2	75	61	61	46	50
3	83	69	68	47	51
4	77	75	65	57	44
5	70	69	62	62	55
6	69	63	60	64	49
7	75	63	60	68	56
8	83	66	61	74	63
9	80	62	62	70	64
11	83	70	60	72	62
12	82	69	65	73	68
13	81	70	63	63	65
Posttest	85	58	70	60	52

One possible explanation for the behavior of these class averages on the various units over time is the fact that teachers in the first semester of a course may tend to teach to the poorer students and then teach to the better students during the second semester. However, achievement profiles run with the upper third of the class on the posttest, the middle third of the class on the posttest, and the lower third of the class on the posttest indicated no such significant trends. Two possible explanations for this phenomenon are (1) that the items were written to distinguish between individuals and not to measure the achievement of the performance criteria and (2) that the items in fact are not measuring the relevant material that was being taught in the class.

References

French, J.W., Ekstrom, R.B., & Price, L.A. Manual for kit of reference tests for cognitive factors. Princeton: Educational Testing Service, 1963.

Gorth, W. & Pinsky, P. Technical memorandum no. TM-18: Demographic, aptitude, and attitude surveys of the students, teachers, and schools in Project CAM. Stanford, Ca.: Project CAM, School of Education, 1968.

Lindeman, R.H., Gorth, W.P., & Allen, D.W. Technical memorandum no. TM-6: The evaluation of item performance in an item sampling case. Stanford, Ca.: Project CAM, School of Education, 1968.

Payne, D.A., & McMorris, R.F. Educational and psychological measurement; contributions to theory and practice. Waltham, Mass.: Blaisdell, 1967.

Pinsky, P., & Gorth, W. Technical memorandum no. TM-17: Monitoring schedules developed for research, 1968-1969. Stanford, Ca: Project CAM, School of Education, 1968.